The University of Sydney

DATA2001: Data Science, Big Data and Data Variety

Practical Assignment

Sydney Liveability Analysis

Authors:

Sharon Davids	460365767
Kimaya Jambhale	510472357
Rubaina Mehrab	500599576

24th May, 2022

Dataset Description

Data Sources

The liveability scores were computed using 7 different datasets from 4 different sources, which were then stored in a database. The 'BusinessStats.csv', 'Neighbourhoods.csv' and 'SA2_2016_AUST.zip' datasets were from the ABS census data. The 'break_and_enter.zip' dataset was from the NSW Bureau of Crime Statistics and Research (BOCSAR), and 'school_catchments.zip' were from NSW Department of Education. Additional datasets, 'swimming pools' and 'bicycle parking', were sourced from the City of Sydney Open Data Hub.

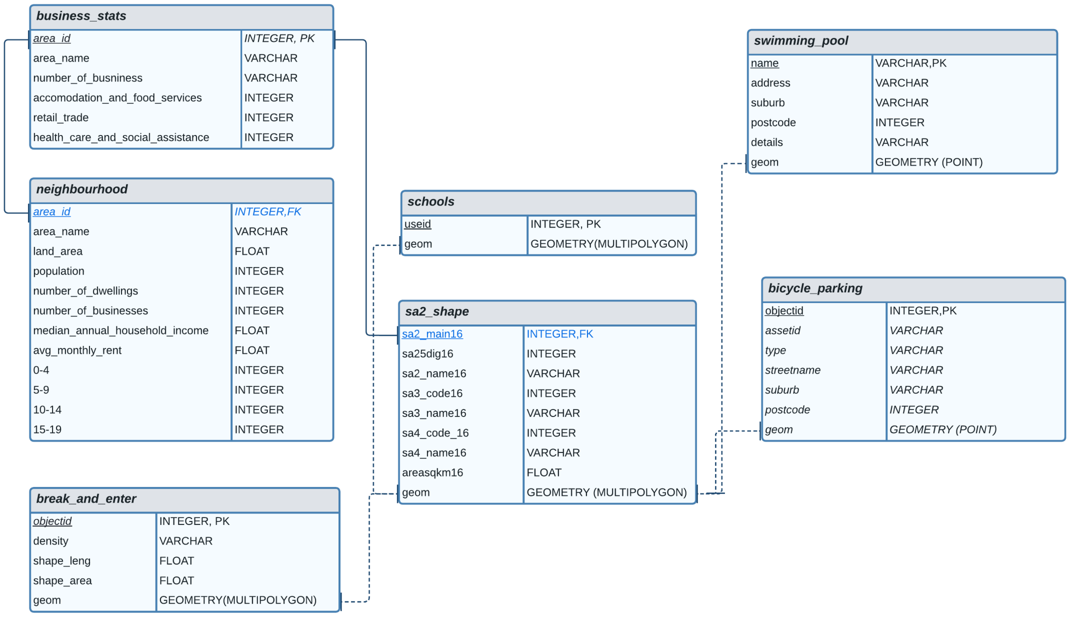
Data Pre-Processing

We worked on copies of the original data to ensure we had a backup in case of any issues. For all datasets, unnecessary columns were removed to speed up queries. We then ensured that the data-types for each attribute was suitable for importing into the database, and made the necessary conversions to achieve this. Duplicate rows were checked and deleted, and an attribute in each dataset was checked for unique values to assign primary keys or foreign keys. Missing cells were also checked and dealt with in ways appropriate to each data attribute. The SRID of geospatial data attributes were found through code for SRID transformations.The following additional steps were done for each dataset:

- **BusinessStats.csv:** 'area_id' was made the primary key seeing as all its values were unique. We also ensured that there were no missing cells.
- **Neighbourhoods.csv:** Missing values for some numerical attributes were filled with 0, as missing data existed in the records for non-residential areas such as industrial areas, business parks, and military areas. Missing cells in the 'population' attribute were filled with 1 to avoid division errors when computing liveability scores.
- **Break_and_enter.zip:** We ensured that the attribute 'geom' was of type 'MultiPolygon', to match geometric data from other datasets, then applied the SRID transformations.
- **School_catchments.zip:** The future catchments data was disregarded as we only considered the present data. Data from primary schools and secondary schools was combined by appending the tables. This resulted in some duplicate rows, as some schools provide both primary and secondary education, which were removed. The geodata of these schools was combined using a union join, and the type was converted to 'MultiPolygon'.
- **SA2_2016_AUST.zip:** Necessary data, i.e. records with 'GCC_NAME16' = 'Greater Sydney' was stored only. It was found that some cells in the 'geometry' attribute had 'None' values, which occurred in records from non-residential areas, hence they were removed. Values for this attribute were either of types 'Polygon' or 'MultiPolygon', hence they were all converted to MultiPolygons for consistency.
- **Swimming pools and Bicycle parking:** The dataset for swimming pools had no missing cells, while that for bicycle parking had two rows with missing values for the 'suburb', 'postcode' and 'streetname' attributes, which were not useful and hence deleted these rows. The geodata from both datasets was of type 'Point', hence no conversion was needed before applying the SRID transformation.

Database Description

Schema



business_stats	
area_id	INTEGER, PK
area_name	VARCHAR
number_of_busniness	VARCHAR
accomodation_and_food_services	INTEGER
retail_trade	INTEGER
health_care_and_social_assistance	INTEGER

neighbourhood	
area_id	INTEGER,FK
area_name	VARCHAR
land_area	FLOAT
population	INTEGER
number_of_dwellings	INTEGER
number_of_businesses	INTEGER
median_annual_household_income	FLOAT
avg_monthly_rent	FLOAT
0-4	INTEGER
5-9	INTEGER
10-14	INTEGER
15-19	INTEGER

break_and_enter	
objectid	INTEGER, PK
density	VARCHAR
shape_leng	FLOAT
shape_area	FLOAT
geom	GEOMETRY(MULTIPOLYGON)

schools	
useid	INTEGER, PK
geom	GEOMETRY(MULTIPOLYGON)

sa2_shape	
sa2_main16	INTEGER,FK
sa25dig16	INTEGER
sa2_name16	VARCHAR
sa3_code16	INTEGER
sa3_name16	VARCHAR
sa4_code_16	INTEGER
sa4_name16	VARCHAR
areasqkm16	FLOAT
geom	GEOMETRY (MULTIPOLYGON)

swimming_pool	
name	VARCHAR,PK
address	VARCHAR
suburb	VARCHAR
postcode	INTEGER
details	VARCHAR
geom	GEOMETRY (POINT)

bicycle_parking	
objectid	INTEGER,PK
assetid	VARCHAR
type	VARCHAR
streetname	VARCHAR
suburb	VARCHAR
postcode	INTEGER
geom	GEOMETRY (POINT)

Figure 1: Database Schema. 'PK' and 'FK' represent the primary and foreign keys, respectively.

It was observed that the 'area_id' attribute of business_stats matched with 'area_id' of the neighbourhood table and 'sa2_main16' column from sa2_shape data. From these tables, the PRIMARY KEY was the 'area_id' of the business_stats table which mapped to neighbourhood and sa2_shape tables as FOREIGN KEYS. The reason why we chose to keep the 'area_id' of the business_stats table as the PRIMARY KEY was because it had the highest number of observations. So we prioritised this dataset over the others.The links among the PRIMARY and FOREIGN keys are shown by the solid lines.

The schools and break_and_enter had no common entries that could be matched with other tables and hence they have their own PRIMARY KEYS which is the 'use_id' and 'objectid' respectively. The other 2 datasets that we imported - swimming_pools and bicycle_parking have their unique PRIMARY KEYS as well which are 'name' and 'objectid'. The databse schema shows the entries we used to join the tables in the form of dashed lines.

Indexes

We created index in our geometry data since most of our queries required spatial joins and as they take a lot of time, we figured creating indexes on them will improve speed and efficiency of the SQL querying. We created 4 indexes in total and they are all on the 'geom' field of the 'sa2_shape', 'break_and_enter', 'schools', 'swimming_pool' and 'bicycle_parking' tables.

Greater Sydney Score Analysis

The liveability score of Greater Sydney was calculated using the supplied formula.

$$Score = \mathcal{S}(z_{school} + z_{accom} + z_{retail} - z_{crime} + z_{health}) \tag{1}$$

To compute each of the z-scores (z_{school}, z_{accom}, z_{retail}. z_{crime}, z_{health}), an SQL query was used to obtain data. The measures and each z-score was calculated in Python. z_{school} was calculated only using the primary and secondary data since we were interested present data only, and this was filtered during the data pre-processing stages. Each SQL query used the 'SA2_shape' data to filter the suburbs or geometry to only Greater Sydney region. Due to the geospatial datasets having different spatial reference identifiers (SRIDs), the function 'ST_Transform' (to 4326/ World Geodetic System) was used in geometry joins to ensure all geodata was using the same SRID, namely '4326' as it was the most commonly used between the 7 datasets.

Depending on the query, different join types were used (inner, left outer, etc.) and in some queries, grouping by sa2_area was done to ensure all data output had the exact number of Greater Sydney SA2 areas (312 rows). All queries were also ordered (using 'order by') by sa2_shape id (area ids) to ensure all tables were in the same order, for score calculations.

After each measure was computed for each SA2 region, the following formula was used to standardise variables and produce the z-scores.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Each of the parameters in the formula for the liveability score has a different range for which values can exist, for instance the 'school' measure is very different in the values it takes compared to the 'accomm' measure which can have much higher values and hence skew the score disproportionately. Thus, computing the z-scores allows for simpler comparisons between each measure considered in the overall liveability of SA2 regions.

To allow for easier comparison between the neighbourhoods, the sigmoid function was applied sum of the parameters. This ensured that all values were between 0 and 1. The sigmoid (or logistic) function is:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The liveability scores of each SA2 area were plotted and overlaid over a map for easy visualisation of the comparative scores of neighbourhoods across Greater Sydney . This is viewed in Figure 2.

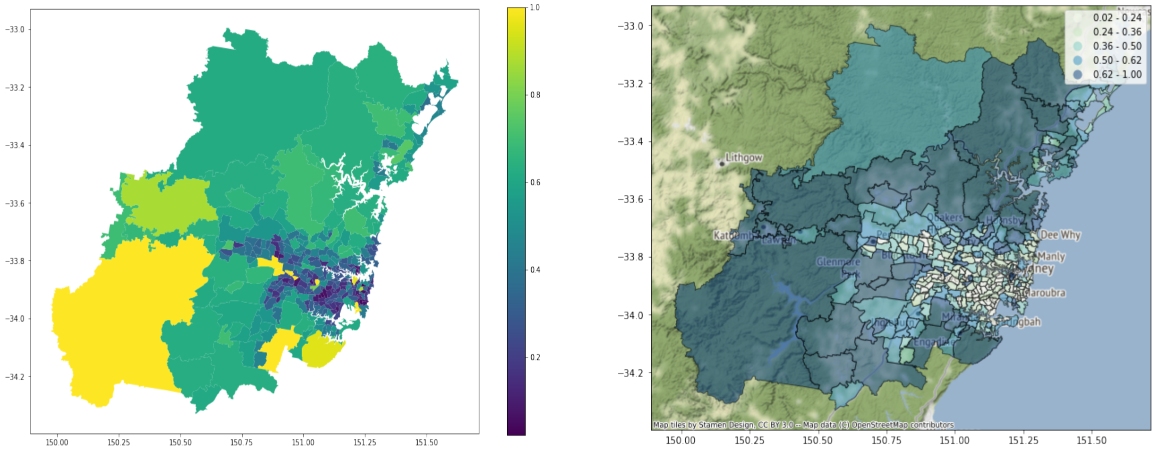


Figure 2: Greater Sydney Score Maps

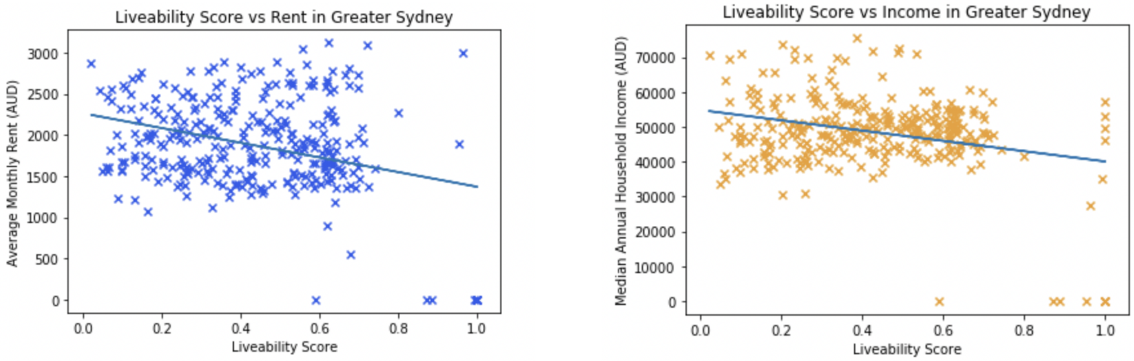
Overview of Results

As seen in Figure 2, SA2 areas closer to Sydney CBD generally have lower liveability scores, while those in the outskirts of the Greater Sydney area, such as the Blue Mountains region, have higher scores. This can be attributed to the way that the liveability score formula was defined, where neighbourhoods closer to Inner Sydney are more populated, meaning that despite higher numbers of school catchment areas, accommodation, food services, retail services and health services, the densities were lower in metropolitan areas, which lowered the sum of z-scores. Due to the relatively low population in Outer Sydney SA2 regions, the total number of crime hotspot areas are also low. Additionally, these regions are much larger in area, which can also be seen graphically. Thus, the density of hotspot areas is quite low, meaning the score for these areas is not decreased by much (see formula).

There are some exceptions to this, however, where some neighbourhoods are more ’liveable’ than those surrounding. This may be because there exist some relatively affluent areas with larger, more spacious homes and in which populations are much lower. This results in lower densities of school catchment areas, retail and health services, such as in neighbourhoods Bella Vista Waters and Northmead, which have access to Norwest Private Hospital and Westmead Hospital, respectively.

Correlation Analysis

Hypothesis: The more liveable areas, i.e. those with greater liveability scores, will generally have a higher average rent and higher median annual income.



(a) Mean Monthly Rent vs Liveability Score

(b) Median Annual Income vs Liveability Score

Figure 3: Scatterplots to show correlation

Correlation refers to the extent to which two quantitative variables are linearly associated. To test the degree of correlation, we fit a linear model and measure how appropriate it is for the data.

Assumptions:

- The liveability scores and mean monthly rent prices are independent of one another. We can confirm this as the mean monthly rent was not used as a parameter in the formula which calculated the liveability scores.
- The same applies for the liveability scores and the median annual income.

Mean Monthly Rent

The r value for the linear relationship between the liveability score and average monthly rent is 0.105 (3dp), and the gradient of the line of best fit is negative (approx. -892), hence the correlation coefficient is $r = -\sqrt{0.105} = 0.324$ (3dp). This indicates that there exists a weak, negative, linear association between the two variables, namely liveability score and average rent in Greater Sydney, and thus a weak correlation. From both the small r value and the points on the scatter plot (Figure 3a), it can be concluded that the points seemingly do not follow a particular trend.

Median Annual Income

The r value for the linear relationship between the liveability score and median annual income is 0.0814 (3dp), and the gradient of the line of best fit is also negative here (approx. -14780), hence the correlation coefficient is $r = -\sqrt{0.0814} = 0.285$ (3dp). Thus, there is a weak, negative, linear association between the liveability score and the median annual income. Although the r-value is lower than that for the previous two variables, from Figure 3b, the two variables here are seemingly more strongly associated than those in Figure 3a, however may not be as strongly correlated.

Interpretation of Results

The hypothesis can be rejected, as there is no strong, positive, linear relationship between the liveability scores and median incomes, and with mean rent. This may be due to factors such as crime, which is generally higher in more affluent areas, i.e. those associated with higher income and rent, thus reducing the overall liveability of these areas.

City of Sydney Analysis

Our stakeholder is a young, single individual in their late twenties, who is looking to work full-time in Sydney. They are active, social, and enjoy a busy lifestyle, and their hobbies include swimming and bike riding. They are looking to rent a space and are healthy, meaning the locality of health services is not a key factor in their decision about where to live. They do not have many valuables and are more concerned about the convenience of food services but are not a big shopper (not worried about retail), hence the crime rate in their neighbourhood is not a key factor in their choice of residence. They would prefer not to live near schools, as they do not plan on having children and usually avoid loud noise from schools. The stakeholder is financially independent, and does not mind the rent price, given that their other criteria are fulfilled.

We had redefined the the previous liveability score (Equation 1) to the following formula for our stakeholder.

$$Score = \mathcal{S}(-0.1z_{school} + 0.3z_{accom} + 0.05z_{retail} - 0.15z_{crime} + 0.05z_{health} + 0.2z_{swim} + 0.15z_{bicycle}) \quad (4)$$

Where $\mathcal{S}$ is the sigmoid function, z is the normal z-score (calculated using equations 2 & 3) , and the meaning of each parameter is shown below. We have defined the liveability score such that a higher score indicates a better match for the stakeholder.

Table 1: Meaning of each parameter in City of Sydney liveability score equation.

Parameter	**Meaning**
z_{school}	Relative density of school catchment areas
z_{accom}	Relative availability of accommodation and food services
z_{retail}	Relative availability of retail services
z_{crime}	Relative prevalence of crime hotspot areas
z_{health}	Availability of health services
z_{swim}	Number of swimming pools per SA2 area
$z_{bicycle}$	Number of bicycle parking spaces available per SA2 area

Given the stakeholder's interests in swimming, the normal z-score for the number of swimming pools is weighted relatively heavily. Bicycle parking is not weighted as heavily because, while it's important, if the stakeholder lives by a swimming pool, a bike can be parked anywhere or if necessary (attached to pole). As the stakeholder would like to avoid areas where they would be more likely to experience noise from school children, the coefficient for the z-score for number of school catchment areas was set to -0.1. The stakeholder highly values the availability of accommodation and food services , hence the parameter representing this information, z_{accom}, was scaled highly, with a coefficient of 0.3. The coefficients of z_{retail} and z_{health} were both quite low (0.05), because while higher densities of retail services and health services are convenient, the stakeholder does not favour them as highly as they favour their hobbies. As this individual is willing to accept any rent price, median rent is not a factor incorporated into the formula. A sigmoid function was applied to the weighted sum of the parameters like the Greater Sydney Liveability Score.

The liveability scores of each SA2 area were plotted and overlaid over a map for the stakeholder to easily identify those neighbourhoods in the City of Sydney more suited to their needs. This is viewed in Figure 4.

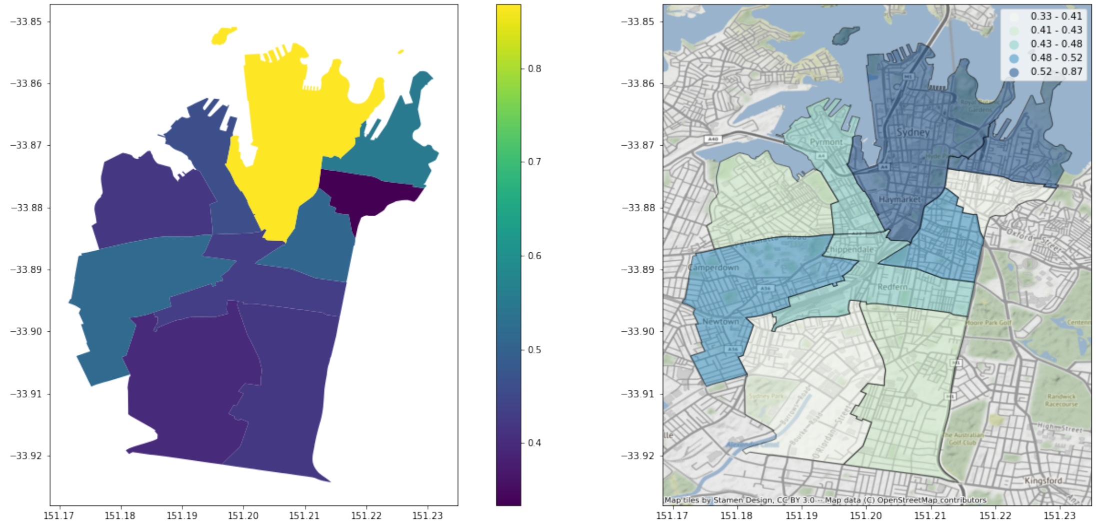


Figure 4: City of Sydney Score Maps

As seen from the second map from Figure 4, the suburbs of Sydney, Haymarket and The Rocks have the highest liveability scores in Sydney Inner City, followed by Potts Point, Woolloomooloo and Newtown and its neighbouring suburbs. Therefore, we would highly suggest that our stakeholder seeks homes in these areas.